参赛密码 _____

(由组委会填写)

“华为杯”第十三届全国研究生数学建模竞赛

学　　校	华东师范大学，上海交通大学
参赛队号	K0247
队员姓名	1.孙凤云
	2.朱云龙
	3.魏宁

参赛密码 _____
（由组委会填写）

题 目 基于假设检验与关联分析的多性状致病位点与致病基因定位方法研究

摘 要：

研究表明人体的许多表型性状差异以及对药物和疾病的易感性等都可能与某些位点相关联，或和包含有多个位点的基因相关联。因此，定位与性状或疾病相关联的位点在染色体或基因中的位置，能帮助研究人员了解性状和一些疾病的遗传机理，也能使人们对致病位点加以干预，防止一些遗传病的发生。针对题中的问题，本文采用如下方法加以解决。

针对第一个问题，我们首先统计每一个位点两个碱基出现的频率，频率多的为主脱氧核苷酸，频率少的为次脱氧核苷酸。若某一位点由两个主脱氧核苷酸组成，则编码为 2，若位点是杂合的，即由一个主脱氧核苷酸和一个次脱氧核苷酸组成，则该位点编码为 1，若位点由两个次脱氧核苷酸组成，则该位点编码为 0。

针对第二个问题，我们采用统计学方法加以解决，本文采用了三种假设检验的方法对9445个位点进行了分析，三种方法分别为费舍尔精确检验、logistic回归分析和Cochran-Armitage trend 检验方法。三种方法均表明"rs2273298"位点的致病可能性最大，"rs2250358"位点致病可能次之，"rs12036216"位点、"rs4391636"位点、"rs7368252"位点、"rs932372"位点、"rs7543405"位点、"rs9426306"位点与"rs12145450"位点也有一定的致病可能性。

针对第三个问题，本文采用费舍尔方法对致病基因进行估计，分别对由费舍

尔精确检验得到的P-value与logistic回归方法得到的P-value进行重要性分析，二者均得到四个致病基因，且其中有三个是重合的。由于费舍尔方法没有考虑位点之间的依赖关系，我们又采用了布朗近似对数据进行了分析，得到的致病基因与费舍尔方法得到的大体一致，但可减小false positive rate。最终我们得到102号基因致病可能性最大，55、217、265、293号基因也有致病可能性。

针对第四个问题，我们采用改进关联分析方法对两个集合的关联性进行了分析。传统的关联分析（canonical correlation analysis）只能解决线性关联。为此我们使用了惩罚典型相关分析，并且通过最优缩放将位点的离散数值编码转换为连续变量，便于进行关联计算。结果显示“rs351617”位点的致病可能性最大，“rs7538876”位点和“rs4949516”位点致病可能性次之，“rs780983”位点、“rs12746773”位点、“rs406985”位点和“rs716325”位点也有致病的可能性。另外还有23个位点存在稍弱的致病可能性。

目　　录

1 问题重述

1.1 问题背景

人体的每条染色体携带一个 DNA 分子，人的遗传密码由人体中的 DNA 携带。DNA 是由分别带有 A，T，C，G 四种碱基的脱氧核苷酸链接组成的双螺旋长链分子。在这条双螺旋的长链中，共有约 30 亿个碱基对，而基因则是 DNA 长链中有遗传效应的一些片段。在组成 DNA 的数量浩瀚的碱基对（或对应的脱氧核苷酸）中，有一些特定位置的单个核苷酸经常发生变异引起 DNA 的多态性，我们称之为位点。在 DNA 长链中，位点个数约为碱基对个数的 1/1000。由于位点在 DNA 长链中出现频繁，多态性丰富，近年来成为人们研究 DNA 遗传信息的重要载体，被称为人类研究遗传学的第三类遗传标记。大量研究表明，人体的许多表型性状差异以及对药物和疾病的易感性等都可能与某些位点相关联，或和包含有多个位点的基因相关联。因此，定位与性状或疾病相关联的位点在染色体或基因中的位置，能帮助研究人员了解性状和一些疾病的遗传机理，也能使人们对致病位点加以干预，防止一些遗传病的发生。

近年来，研究人员大都采用全基因组的方法来确定致病位点或致病基因，具体做法是：招募大量志愿者（样本），包括具有某种遗传病的人和健康的人，通常用 1 表示病人，0 表示健康者。对每个样本，采用碱基（A，T，C，G）的编码方式来获取每个位点的信息（因为染色体具有双螺旋结构，所以用两个碱基的组合表示一个位点的信息）；例如本题中在位点 rs100015 位置，不同样本的编码都是 T 和 C 的组合，有三种不同编码方式 TT，TC 和 CC。类似地其他的位点虽然碱基的组合不同，但也只有三种不同编码。研究人员可以通过对样本的健康状况和位点编码的对比分析来确定致病位点，从而发现遗传病或性状的遗传机理。

1.2 问题分析

为了通过全基因组的方法来确定致病位点或致病基因，本题提出了下面四个问题，下面我们对问题进行重述，并简要叙述解决方案。

问题一、请用适当的方法，把 genotype.dat 中每个位点的碱基（A，T，C，G）编码方式转化成数值编码方式，便于进行数据分析。

解决方案：通过对每个位点所在列进行观察，我们发现每一列只有三种编码方式，因此我们可以采用 0/1/2 的方式对每一个位点进行编码，编码时我们考虑到了次脱氧核苷酸（编码为 0）和主脱氧核苷酸（编码为 2）的问题，为后面的分析提供了方便。

问题二、根据附录中 1000 个样本在某条有可能致病的染色体片段上的 9445 个位点的编码信息 (见 genotype.dat) 和样本患有遗传疾病 A 的信息（见 phenotype.txt 文件）。设计或采用一个方法，找出某种疾病最有可能的一个或几个致病位点，并给出相关的理论依据。

解决方案：这里我们通过构造假设检验的方式来找出可能的致病位点，特别的为了保证所得结果的可靠性，我们采用费舍尔精确检验、Cochran-Armitage trend 检验、logistics 回归中 β 是否为 0，这三种检验得到的 p 值进行分析，同时我们也考虑到了多重假设检验所带来的问题，对所得的 p 值进行了修正。

问题三、同上题中的样本患有遗传疾病 A 的信息（phenotype.txt 文件）。现有 300 个基因，每个基因所包含的位点名称见文件夹 geneinfo 中的 300 个 dat 文件，每个 dat 文件列出了对应基因所包含的位点 (位点信息见文件 genotype.dat)。由于可以把基因理解为若干个位点组成的集合，遗传疾病与基因的关联性可以由基因中包含的位点的全集或其子集合表现出来请找出与疾病最有可能相关的一个或几个基因，并说明理由。

解决方案：因为每个基因含有多个位点，为了充分利用每一个位点的信息，我们通过费舍尔方法，分别对第二问中三种假设检验得到的 p 值进行融合，通过融合后得到的检验统计量对每一个基因与疾病的关联性进行假设检验，找出可能的致病基因。

问题四、在问题二中，已知 9445 个位点，其编码信息见 genotype.dat 文件。在实际的研究中，科研人员往往把相关的性状或疾病看成一个整体，然后来探寻与它们相关的位点或基因。试根据 multiphenos.txt 文件给出的 1000 个样本的 10 个相关联性状的信息及其 9445 个位点的编码信息 (见 genotype.dat)，找出与 multiphenos.txt 中 10 个性状有关联的位点。

解决方案：我们使用了惩罚典型相关分析，并且通过最优缩放将位点的离散数值编码转换为连续变量，便于进行关联计算。结果显示“rs351617”位点的致病可能性最大，“rs7538876”位点和“rs4949516”位点致病可能性次之，“rs780983”位点、“rs12746773”位点、“rs406985”位点和“rs716325”位点也有致病的可能性。

2　模型假设与符号说明

2.1　符号说明

表 1　符号说明

I	样本数量
J	位点数量
D_1	主脱氧核苷酸
D_2	次脱氧核苷酸
s_{ij}	第 i 个样本第 j 个位点的数值编码
P-value	假定值
p_j	位点 j 的假定值
X_k^2	自由度为 k 服从 X^2 的假设检验
g_{jk}	第 k 个基因是否包含第 j 个位点
t_j	第 j 个位点的自然底数值的二倍
χ_m^2	自由度为 m 的卡方分布
σ^2	第 k 个基因对应的位点协方差之和
Ω	t 的协方差矩阵
ω_i^T	第 i 个特征的权重向量
$\rho_{u,v}$	u 和 v 的 Pearson 相关系数
Σ_{ij}	第 i 个向量与第 j 个向量的协方差
λ	拉格朗日系数
θ	拉格朗日系数
$\hat{a}^k$	第 k 次迭代所估计的 X 的权重向量
$\hat{b}^k$	第 k 次迭代所估计的 Y 的权重向量
x^*	转换后得到的连续型变量
$\mathfrak{J}$	转换函数
G_j	指示矩阵
c_j	变量 j 的分类量化结果

3　问题一模型建立与求解

DNA 即脱氧核糖核酸，是一种生物大分子，可组成遗传指令，引导生物发育与生命机能运作。主要功能是信息储存，可比喻为“蓝图”或“食谱”。其中包含的指令，是建构细胞内其他的化合物，如蛋白质与核糖核酸所需。带有蛋白质编码的 DNA 片段称为基因。DNA 是一种长链聚合物，组成单位为四种脱氧核苷酸，即：腺嘌呤脱氧核苷酸 (A) 、胸腺嘧啶脱氧核苷酸 (T)、胞嘧啶脱氧核苷酸 (C)、鸟嘌呤脱氧核苷酸 (G)。四种碱基两两配对，即（A-T，C-G 相互作用）形成 DNA 单体以及编码遗传信息的化学结构。

在组成 DNA 的数量浩瀚的碱基对（或对应的脱氧核苷酸）中，有一些特定位置的单个核苷酸经常发生变异引起 DNA 的多态性，我们称之为位点。在 DNA 长链中，位点个数约为碱基对个数的 1/1000。由于位点在 DNA 长链中出现频繁，多态性丰富，近年来成为人们研究 DNA 遗传信息的重要载体，被称为人类研究遗传学的第三类遗传标记。

通过对 1000 个样本的位点信息的数据分析，我们可知，每个位点都包含两个碱基对，即可用两个脱氧核苷酸表示，并且其编码方式有三种，即两个碱基对的排列方式为 2*2-1。如位点“rs3094315”可用 C 与 T 的组合方式表示，其三种编码为 CC、TT 与 TC。通过对每一个位点 1000 个样本的两种脱氧核苷酸的数量进行统计，我们可得到该位点的主脱氧核苷酸（数量多的脱氧核苷酸）与次脱氧核苷酸（数量少的脱氧核苷酸)。再如“rs3094315”位点，1000 个样本中，TT 有 664 个，TC 有 293 个，CC 有 43 个，因此 T 的数量为 1621 个，C 的数量为 379 个，所以在该位点中 T 为主脱氧核苷酸，C 为次脱氧核苷酸。文献 [1] 中对基因型的编码方式，本文对位点碱基编码方式进行如下的数值编码方式的转换。

假定数据为从 I 个个体中采集的 J 个位点的信息，并且 J 远远大于 I，即 $J \gg I$，本文中，$I = 1000$，$J = 9445$。假设第 j 个位点包含的两种碱基分别为 D_1，D_2，其中 D_1 为主脱氧核苷酸，D_2 为次脱氧核苷酸。则第 i 个个体的第 j 个位点的数值 s_{ij} 如公式（1）所示

$$s_{ij} = \begin{cases} 0, & \text{if } D_2D_2 \\ 1, & \text{if } D_1D_2 \text{or } D_2D_1 \\ 2, & \text{if } D_1D_1. \end{cases} \tag{1}$$

也即如果该位点包含两个主脱氧核苷酸则该位点编码为 2，如果该位点为杂合的（包含一个主脱氧核苷酸和一个次脱氧核苷酸)，该位点编码为 1，如果该位点包含两个次脱氧核苷酸，该位点编码为 0。因此所有位点均为 0，1，2 三种数值码。一段数值编码的示意图如图 1 所示。

rs3094315	rs3131972	rs3131969	rs1048488	rs1256203	rs1212481	rs4040617	rs2980300	rs4970383	rs4475691	rs1806509	rs7537756	rs2340587	rs2857669	rs1110052	rs3748592
2	1	2	1	0	2	2	2	2	2	2	2	1	0	2	2
1	1	1	1	2	1	2	2	2	2	2	2	2	2	2	2
2	0	2	0	2	1	0	2	2	2	2	2	1	0	2	2
2	2	2	1	1	2	1	1	1	1	1	2	2	2	2	2
1	1	1	2	1	2	1	2	0	2	1	2	2	1	1	2
1	2	2	1	0	2	2	2	0	0	2	1	1	1	0	1
1	1	1	1	1	1	1	2	1	1	2	2	2	1	2	2
2	0	1	1	1	2	1	1	1	2	1	1	1	1	2	2
2	2	1	1	2	2	2	0	0	0	1	1	1	0	0	2
2	1	2	1	2	1	2	1	2	2	1	2	2	2	1	2
1	2	2	2	1	1	1	2	2	1	1	2	2	2	1	1
0	0	2	1	1	2	2	2	1	1	1	1	0	0	1	1

图 1　数值编码示意图

Number of	Not rejected	Rejected	
H_0 True	U	V	M_0
H_0 False	T	S	M_1
Total	$M-R$	R	M

图 2　多重检验各类情况图列联表

4　问题二模型建立与求解

我们收集到了 1000 个样本中患 A 病的情况，其中 500 人为 A 病患组，另 500 人为对照组。本题要根据 1000 个样本的患病情况以及与其对应的 9445 个位点信息确定一个或几个可能致病的位点。为解决此问题，本文采用了几种基于假设检验 p 值的方法，这里首先介绍一些假设检验的基本知识。

4.1　假设检验概述

设有 M 个假设检验，$H_1, H_2, ..., H_M$，相应的 p 值为 $p_1, p_2, ..., p_M$，通常情况下，如果 p 值小于 0.05，我们称这个假设检验是显著的。然而此处我们的问题是多重假设检验问题以 0.05 作为 p 值的比较阈值是不合适的。具体原因如下：以 0.05 为阈值时，对于 M 个假设检验，所有检验都正确的概率是 $(1-0.05)^M$，即我们至少犯一个错误的概率为 $1-(1-0.05)^M$，当 M 变大时该概率会很大，在我们的问题中 M 为 9445，该值近似为 1。另一方面，由大数定律可知，在 $M = 9445$ 阈值为 0.05 条件下，平均意义下应该有 $9445 * 0.05 \approx 472$ 个检验被拒绝。

因此我们应对 p 值进行修正，一种方法采用比较常用的方法的 Bonferroni 修正，即将阈值改为 $0.05/M = 0.05/9445 \approx 5.3 \times 10^{-6}$. 另一种是控制错误发现率 FDR,

$$\begin{cases} E(V/R), & \text{if } R > 0 \\ 0, & \text{if } R = 0. \end{cases} \tag{2}$$

其中 V, R 分别为 M 个假设检验中被错误拒绝的检验的个数和拒绝的总个数，如图 2 所示： 我们可以通过 Benjamini-Hochberg 过程来达到控制 FDR 的目的：

- 固定错误发现率，如 $\alpha = 0.05$, 对 p 值排序 $p_{(1)} \le p_{(2)} \le ... \le p_{(M)}$
- $$L = \max\{j : p_{(j)} < \alpha \frac{j}{M}\}$$
- 拒绝所有满足 $p_{(j)} \le p_{(L)}$ 的 H_j.

通过上面两种修正，我们便可以解决多重假设检验的问题。下面我们将分别采用费舍尔精确检验、Cochran-Armitage trend 检验、logistics 回归中 β 是否为 0，这三种检验得到的 p 值进行分析。

4.2　费舍尔精确检验

费舍尔精确检验 [2] 是一种通过分析列联表来判断统计显著性的检验方法。尽管经常在样本数量较小的情况下应用，但它在任意样本数量下都是有效的。它是以它的发明者 Ronald Fisher 命名的，是精确检验方法的一种。之所以被称为精确检验是因为原

表 2　某一统计量的示例列联表

	$Class_1$	$Class_2$	Row Total
Case	a	b	a+b
Control	c	d	c+d
Column Total	a+c	b+d	a+b+c+d=n

表 3　示例 2×3 列联表

	$Class_1$	$Class_2$	$Class_3$	Row Total
Case	a	b	c	a+b+c
Control	d	e	f	d+e+f
Column Total	a+d	b+e	c+f	a+b+c+d+e+f=n

假设的偏差显著性（P-value）能够被准确计算出而不像其他统计实验方法依赖于样本量的增大来近似给出假定值。

费舍尔检验多数情况下被用作对 2×2 的列联表的检验。如果表的边缘分布（每类的总数）是确定的，P-value 能够从该检验中计算得到。当样本量比较大的时候，卡方检验可以被用来处理 P-value 估计，然而它给出的仅是 P-value 的估计值。因为用来计算检验的样本分布仅是理论卡方分布的近似，当样本量较小或者表中元素相差较大时，这种近似是不准确的。

假设某一统计量的列联表如表2 所示，其中行表示病例与对照。其 P-value 是由公式（3）的超几何分布给出。

$$p = \frac{\begin{pmatrix} a+b \\ a \end{pmatrix}\begin{pmatrix} c+d \\ c \end{pmatrix}}{\begin{pmatrix} n \\ a+c \end{pmatrix}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{3}$$

其中 $\begin{pmatrix} n \\ k \end{pmatrix}$ 为二项式系数。

4.3　基于费舍尔精确检验的致病位点分析

如前文所述，费舍尔精确检验主要解决 2×2 列联表的检验问题。而本文所研究的问题为 2×3 列联表的假设检验问题，为此，本文采用针对 2×3 列联表的假设检验问题。具体的对于此问题我们检验的原假设为：列联表的行中有病和无病两种情况关于列联表的列中三种碱基编码方式的分布是相同的。此检验是合理的，因为原假设成立说明有病无病不对对应位点碱基编码分布造成影响。

若某一采样的列联表如表3 所示，则其 P-value 由公式（4）给出。

$$p = \frac{(a+d)!(b+e)!(c+f)!(a+d+c)!(d+e+f)!}{a!b!c!d!e!f!n!} \tag{4}$$

这里我们以“rs3094315”位点为例来说明每个点位的 P-value 计算过程，“rs3094315”位点的列连表如表4所示，其中列分类的“0”、“1”、“2”表示问题一中的数值编码。由公式（4）可得“rs3094315”位点的 P-value 如公式（5）所示。

$$p_{rs3094315} = \frac{43! \times 293! \times 664! \times 500! \times 500!}{26! \times 147! \times 327! \times 17! \times 146! \times 337! \times 1000!} = 0.396 \tag{5}$$

因此可知“rs3094315”位点的 P-value 为 0.396，若以 0.05 为为假设检验的接受概

表 4　“rs3094315”点位列联表

	0	1	2	Row Total
Health	26	147	327	500
Disease	17	146	337	500
Column Total	43	293	664	1000

率，则该位点显然应该被拒绝，即该位点不是致病位点。这与我们的经验分析或直观感受是吻合的。从表4中可以看出，无论从数量还是变化趋势来看，“rs3094315”位点的健康与基本样本的“0”、“1”、“2”编码都比较相似，表明该位点在两类人群中并没有明显的区别，因此该位点并不是致病位点。“rs3094315”位点的健康与疾病人群的编码分布直方图如图 3 所示，从该图中也可看出“rs3094315”位点的健康与疾病人群的编码分布基本一致。

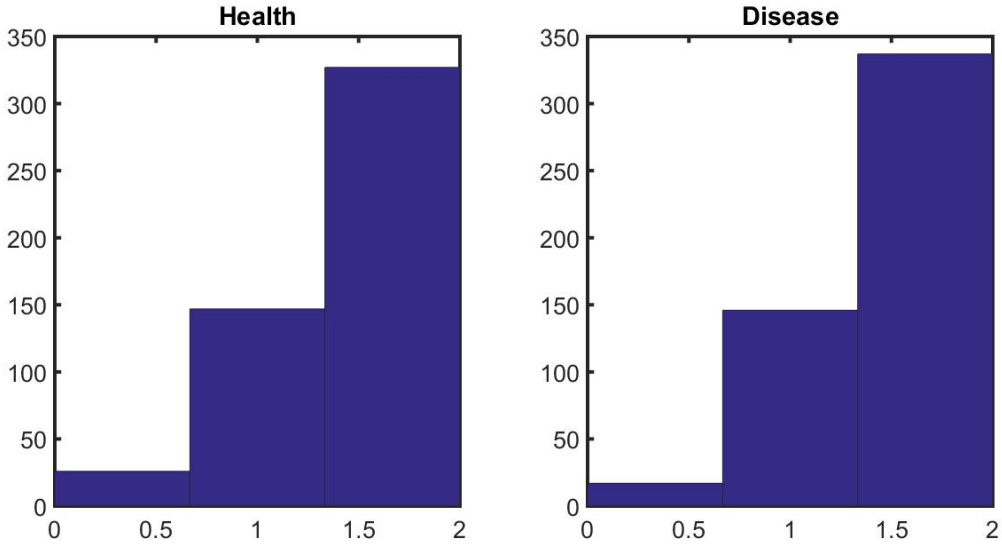

图 3　“rs3094315”位点的健康与疾病人群的编码分布直方图

对所有的 9445 个位点都做以上的分析，便可得到这 9445 个位点的 P-value。在 Bonferroni 修正下，阈值为 $0.05/9445 \approx 5.3 \times 10^{-6}$，比此阈值小的 p 值只有一个 5.474047×10^{-7}，即选出的致病位点只有一个，为“rs2273298”位点，该位点的分布表如表5所示，该位点的分布直方图如图 4 所示。相比于“rs3094315”位点，“rs2273298”位点的健康与疾病人群的编码分布差异较大。本文中也分别给出了选取了阈值为 1.0×10^{-3} 与 1.0×10^{-4} 时的结果，它们选出的致病位点的数量分别为 76 与 9 个。不同阈值与它们选出的位点如表6所示。由此我们可得出结论，“rs2273298”位点的致病嫌疑最大，“rs2250358”位点致病嫌疑次之，“rs12036216”位点、“rs4391636”位点、“rs7368252”位点、“rs932372”位点、“rs7543405”位点、“rs9426306”位点与“rs12145450”位点也有一定的致病嫌疑。进一步的我们采用控制 FDR 在 0.05 水平下的方法进行 BH 修正，修正后最小的 p 值为 0.005170237 为“rs2273298”位点，第二小的 p 值即为 0.336419006，故只有“rs2273298”位点的检验是显著的，得到与 Bonferroni 修正一样的结论，故不对其他位点详细分析。

4.4　Cochran-Armitage trend 检验

Cochran-Armitage trend 也常被称为趋势卡方检验，该方法是检验基因型的 2×3 列联表中一种常用方法，采用该方法的一个好处是可以充分利用第一问的编码方式，那里我们把主脱氧核苷酸对编码为 2，次脱氧核苷酸对编码为 0。而 Cochran-Armitage trend

表 5 “rs2273298”点位列联表

	0	1	2	Row Total
Health	34	161	305	500
Disease	60	218	222	500
Column Total	94	379	527	1000

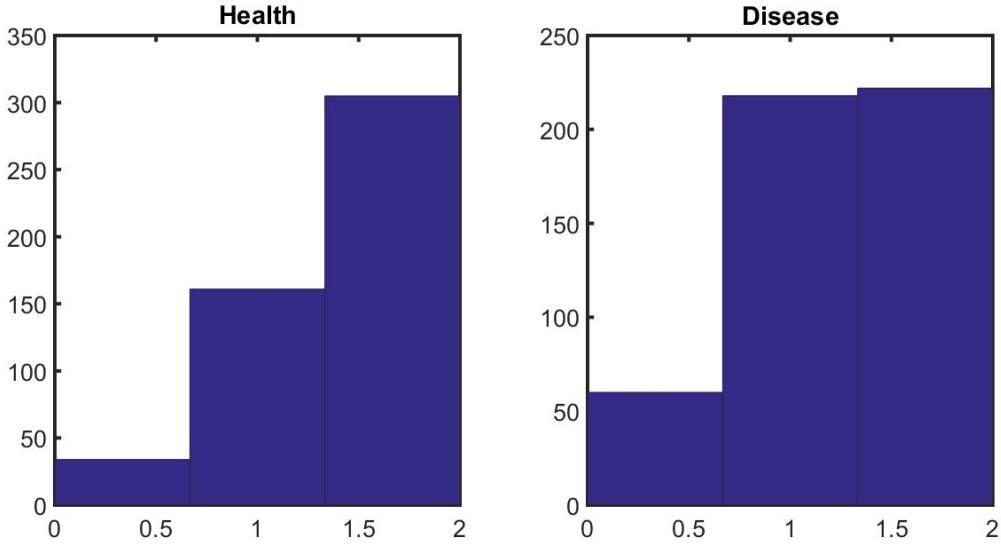

图 4 “rs2273298”位点的健康与疾病人群的编码分布直方图

检验恰好适用于这种编码是有序的情况。具体的对于我们的问题，原假设为列联表的行中有病和无病两种情况关于列联表的列中三种碱基编码的有序趋势是相同的。此检验是合理的，因为原假设成立说明有病无病不对对应位点碱基编码有序趋势造成影响。

下面给出该检验的检验统计量及其满足的分布，对于第 j 个位点 trend 检验统计量可以写成如下形式:

$$T_j = \frac{n_1 + 2n_2 - p(N_1 + 2N_2)}{\sqrt{p(1-p)(N_0N_1 + N_1N_2 + 4N_0N_2)/I}}, \quad j = 1, ..., J \tag{6}$$

其中:

$$p = \frac{\sum_{i=1}^{I} y_i}{I}, y_i = 0, 1 \tag{7}$$

$$n_a = \sum_{i=1}^{I} y_i \delta(s_{ij} = a), \quad a = 1, 2; \tag{8}$$

表 6 费舍尔精确检验得到的不同接受值与其选出的致病位点

阈值	位点数量	位点名称
1.0×10^{-4}	9	“rs12036216”、“rs4391636”、“rs7368252” “rs2250358”、“rs2273298”、“rs932372” “rs7543405”、“rs9426306”、“rs12145450”
1.0×10^{-5}	2	“rs2250358”、“rs2273298”
1.0×10^{-6}	1	“rs2273298”
1.0×10^{-7}	1	“rs2273298”

表 7　Cochran-Armitage trend 检验得到的不同接受值与其选出的致病位点

阈值	位点数量	位点名称
1.0×10^{-4}	4	“rs12036216”、“rs2273298”、“rs932372”、“rs7543405”
1.0×10^{-5}	2	“rs2273298”
1.0×10^{-6}	1	“rs2273298”

表 8　logistic 回归得到的不同接受值与其选出的致病位点

阈值	位点数量	位点名称
1.0×10^{-4}	3	“rs7368252”、“rs2273298”、“rs932372”
1.0×10^{-5}	2	“rs2273298”
1.0×10^{-6}	1	“rs2273298”

$$N_b = \sum_{i=1}^{I} \delta(s_{ij} = b), \quad a = 1, 2; \tag{9}$$

上面的 $\delta(.)$ 为示性函数，当括号内式子成立时取值为 1 否则取值为 0。

原假设成立的条件下该统计量服从正态分布，我们可以由此计算出 genotype.dat 中每一列，即每一个位点相应得 p 值，具体计算采用了 R 语言的 coin 包中的 $independence_{test}$ 函数，算得 p 值，下面表给出比较显著的几个位点并加以分析：

对所有的 9445 个位点都做以上的分析，便可得到这 9445 个位点的 P-value。在 Bonferroni 修正下，阈值为 $0.05/9445 \approx 5.3 \times 10^{-6}$，比此阈值小的 p 值只有一个 1.672333×10^{-7} 即选出的致病位点只有一个，为“rs2273298”位点，与上面基于费舍尔精确检验得到的位点一致。同样上表中也给出了以 1.0×10^{-4} 作为阈值所得到的位点：“rs12036216”、“rs2273298”、“rs932372”、“rs7543405”，对比费舍尔精确检验中得到的位点，发现他们都出现在了费舍尔精确检验所得到的位点中。由此可以进一步确定“rs2273298”的致病性，以及“rs12036216”、“rs932372”、“rs7543405”的可能致病性。进一步的我们采用控制 FDR 在 0.05 水平下的方法进行 B-H 修正，修正后最小的 p 值为 0.001579518 为“rs2273298”位点，第二小的 p 值即为 0.252081234，故只有“rs2273298”位点的检验是显著的，得到与 Bonferroni 修正一样的结论，故不对其他位点详细分析。

4.5　基于 logistic 回归给出的检验

logistic 回归与线性回归类似，只是因变量不再是连续变量而是 0/1 的变量。我么可以把 logistic 回归应用于该问题的每一个位点。具体的：对于样本中的第 i 个观测，因变量 Y_i 取值为 phenotype.txt 中的 0/1，自变量 X_i 为 genotype.dat 中每列编码后的 0/1/2。相应的有 logistic 回归模型如下：

$$logit(P_i) = \beta_0 + \beta_1 X_i \tag{10}$$

其中 $P_i = E(Y_i|X_i)$,$logit(P_i) = \log \frac{P_i}{1-P_i}$. 我们可以对 logistic 回归中的 β_1 是否为 0 进行假设检验，这样检验对于我们的问题是合理的，因为若原假设成立即 $\beta_1 = 0$，则说明 $P_i = E(Y_i|X_i)$ 的值不受位点碱基编码的取值 X_i 的影响。检验 β_1 是否为 0 的方法很多，我们这里直接使用 R 语言中 glm 模型所输出的检验 p 值，下表中给出 p 值比较显著的几个位点并加以分析：

对所有的 9445 个位点都做以上的分析，便可得到这 9445 个位点的 P-value。在 Bonferroni 修正下，阈值为 $0.05/9445 \approx 5.3 \times 10^{-6}$，比此阈值小的 p 值只有一个

2.265398×10^{-7} 即选出的致病位点只有一个，为“rs2273298”位点，与上面基于费舍尔精确检验，Cochran-Armitage trend 检验得到的位点一致。同样上表中也给出了以 1.0×10^{-4} 作为阈值所得到的位点：“rs7368252”、“rs2273298”、“rs932372”，对比费舍尔精确检验中得到的位点，发现他们都出现在了费舍尔精确检验所得到的位点中，并且与 Cochran-Armitage trend 检验得到的位点也只有一个是不同的。由此可以进一步确定“rs2273298”的致病性，以及“rs7368252”、“rs932372”的可能致病性进一步的我们采用控制 FDR 在 0.05 水平下的方法进行 B-H 修正，修正后最小的 p 值为 0.002139668 为“rs2273298”位点，第二小的 p 值即为 0.301113036，故只有“rs2273298”位点的检验是显著的，得到与 Bonferroni 修正一样的结论，故不对其他位点详细分析。

4.6 本章小结

上面我们分别采用费舍尔精确检验、Cochran-Armitage trend 检验、logistics 回归中 β 是否为 0，这三种检验对 9445 个位点分别进行了分析，根据每个位点 P-value 值的大小来判定该位点是否为致病位点, 特别的我们考虑了多重检验问题，并采用了 Bonferroni 修正和 B-H 过程对 p 值进行了修正。三种检验下都选出了的致病位点为”rs2273298”，如图 5 所示： 图中横坐标为 9445 个位点在 genotype.dat 中顺

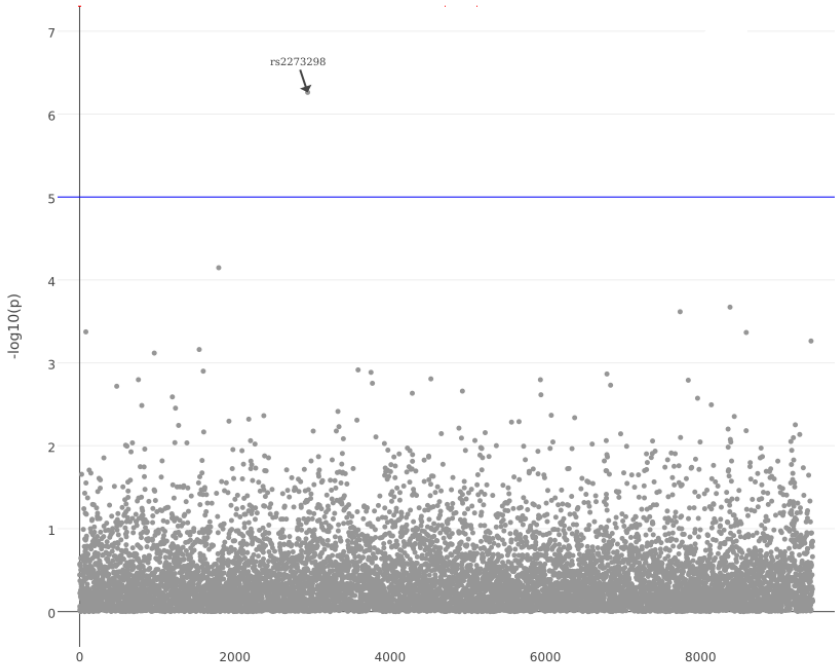

图 5　9445 个位点 P-value 以 10 为底的对数图

序，纵坐标为费舍尔精确检验中 p 值得变换 $-log10(p)$, 蓝线为纵坐标 5，与阈值中 10^{-5} 对应，可以明显的看到，阈值以上的只有标注出来的“rs2273298”这一个位点。除了“rs2273298”位点以外，“rs932372”位点在三种方法下都是相对显著的，“rs12036216”、“rs7543405”、“rs7368252”都在两种方法下相对显著，因此我们有足够的理由相信这四个位点可能为致病位点。最后“rs4391636”、“rs7368252”、“s2250358”、“rs9426306”、“rs12145450”在费舍尔精确检验下相对显著，它们也值得进一步研究。

5 问题三模型建立与求解

现有 300 个基因并且已知它们所包含的位点，利用问题二得到的每个位点与致病性之间的关系，需要找出 300 个基因中可能的致病基因。这里我们基于问题二中三种检验方法得到的每个位点的 P-value，通过费舍尔方法对相应 p 值进行融合，找出致病概率较大的基因。

5.1 费舍尔方法

统计学中，费舍尔方法 (Fisher's method)[3, 4]，也被称作费舍尔联合概率检验是一种用来做数据融合或元分析 (meta-analysis) 的方法。可以利用该方法将同一原假设 H_0 的多次独立检验结果结合到一起，对多次检验进行整体的检验。

费舍尔方法结合每次检验的 P-value，把它们以公式（11）统一到一个检验统计（X^2）中。

$$X_{2k}^2 \sim -2\sum_{i=1}^{k} ln(p_i) \tag{11}$$

其中 p_i 为第 i 次假设检验的 P-value。当 P-value 较小时，检验统计 X^2 较大，表明对于每一次检验原假设都被拒绝。当所有的原假设都为真并且 p_i（或者它们对应的检验统计）是独立的，X^2 是以 $2k$ 为自由度的卡方分布，其中 k 为被联合的检验次数。这一性质能够被用来决定 X^2 的 P-value。

下面解释下为什么检验统计量 X^2 的分布是卡方分布。对于第 i 次检验在原假设成立的条件下，P-value p_i 服从 $[0, 1]$ 上的均匀分布。均匀分布的负自然对数服从指数分布，指数分布乘以 2 后服从自由度为 2 的卡方分布。最后，k 个自由度为 2 的独立卡方分布的和服从自由度为 $2k$ 的卡方分布。有了该检验统计量的分布，我们便可以基于它进行假设检验。

5.2 基于费舍尔方法的致病基因定位

首先定义一个指示函数 g_{jk}，其定义如公式（12）所示。并且自由度为 2 的卡方分布为 $t_j = -2\ln p_j$。基于费舍尔方法，每一个基因的显著性可有公式（13）计算得到

$$g_{jk} = \begin{cases} 1, & if \quad the \quad jth \quad SNP \quad is \quad in \quad the \quad ith \quad gene \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

$$\sum_{j=1}^{9445} g_{jk} t_j \sim \chi_m^2 \tag{13}$$

其中 $m = 2\sum_{j=1}^{9445} g_{jk}$。得到每个位点的显著性信息之后，我们用卡方分布对其进行检验。

首先，给出融合费舍尔精确检验得到的 p 值，其结果如图 6，为了便于观察，我们对 p 进行 $-\log_{10} p$ 变换，并且纵坐标以 4 为界画线。由图中可知，该方法选择的可能致病的基因为 102、217、265 与 293。其中第 102 号基因中包含 10 个位点，且其中有问题二中选出的致病位点“rs2273298”，而其他位点基因包含的位点较多，第 217 号基因包含 20 个位点，第 265 与第 293 号基因分别包含 56 与 60 个位点，且这三个基因中都不包含“rs2273298”位点。因此，本文考虑第 102 号是最有可能致病的基因。 接下来我们融合 Cochran-Armitage trend 检验的到的 p 值，结果如图 7 所示，所选出来的四个致病基因分别为 55、102、217、293，与费舍尔精确检验得到的 P-value 计算出来的致病基因基本吻合。

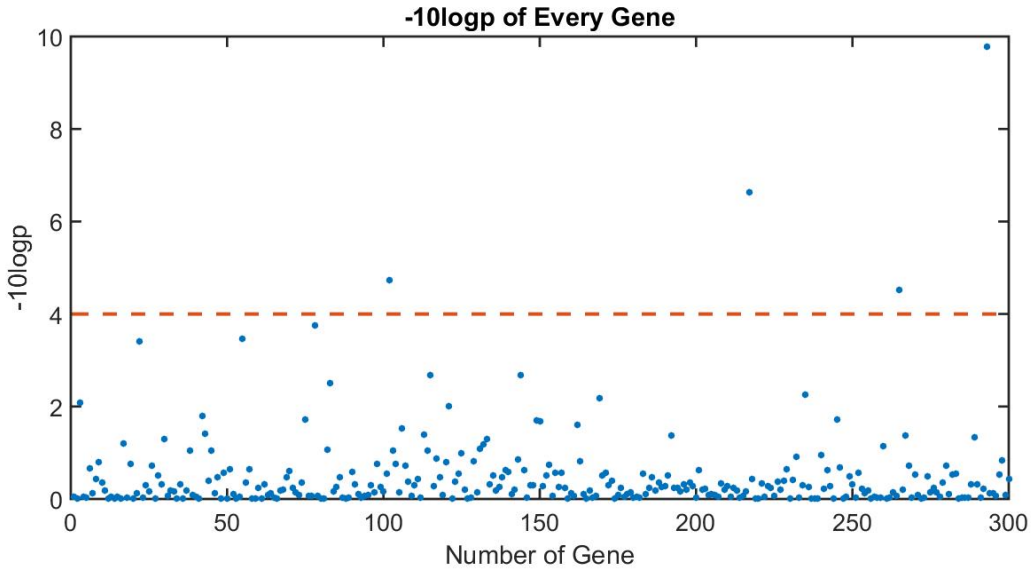

图 6　费舍尔精确检验得到的 300 个基因的 $-10\log p$

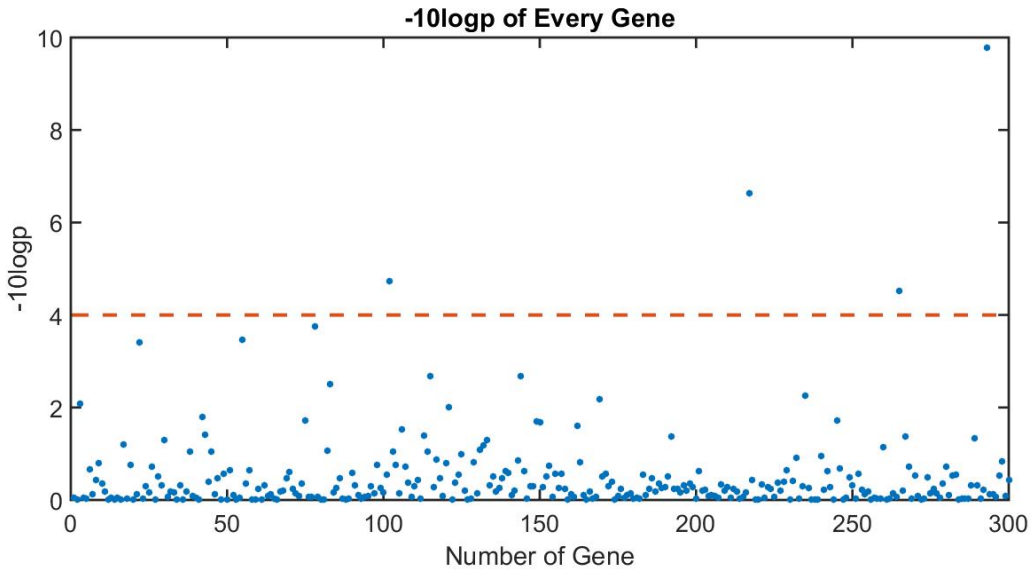

图 7　Cochran-Armitage trend 检验方法得到的 300 个基因的 $-10\log p$

此外，本文也采用了用 logistic 回归方法进行了 p 值计算，并把它们用上述的费舍尔方法计算统计重要性，结果如图 8 所示，所选出来的四个致病基因分别为 55、102、217、293，与费舍尔精确检验得到的 P-value 计算出来的致病基因基本吻合。

5.3　基于布朗近似的致病基因的定位

尽管费舍尔方法非常简洁，并且 Little 与 Folks[5] 证明在检验的数量增加的情况下，费舍尔方法是渐进 Bahadur 最优的。然而位点间的独立假设并不能被保证，因此基于独立假设的基因重要性分析的效果会大打折扣，出现 false positive 结果。因此本文中尝试了布朗近似（Brown's approximation）方法 [6]，如公式（14）所示。

$$\frac{2m\sum_{j=1}^{9445} g_{jk}t_j}{\sigma^2} \sim \chi^2_{2m^2/\sigma^2} \tag{14}$$

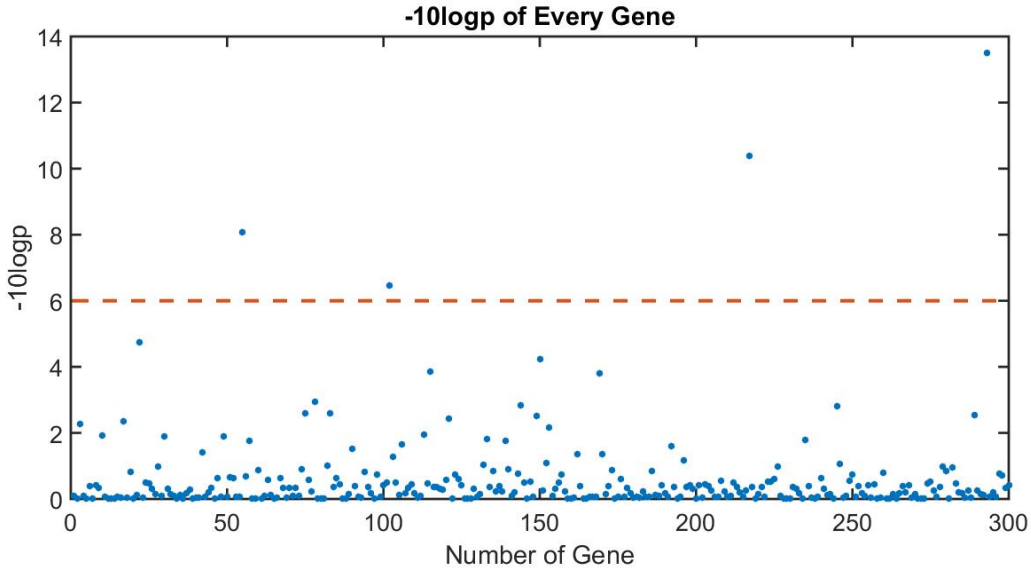

图 8　以 logstic 方法得到的 P-value 计算的 300 个基因的 $-10\log p$

其中 $\sigma^2 = g_{jk}^T \Omega g_{jk}$，$\Omega$ 为 $t = [t_1, t_2, \ldots, t_{300}]$ 的协方差矩阵。Brown 在他的文章中指出在 t_j 不是高度负相关时，布朗近似可取得较好效果。因为此处协方差矩阵 Ω 是未知的，我们需要对它进行估计。我们选择对 1000 个样本的健康与否的二值进行洗牌的方式估计 t 的协方差矩阵，本文中洗牌的次数为 100 次。用布朗近似得到的图如图 9 所示，从图中可知，该方法得到的致病基因为 102。

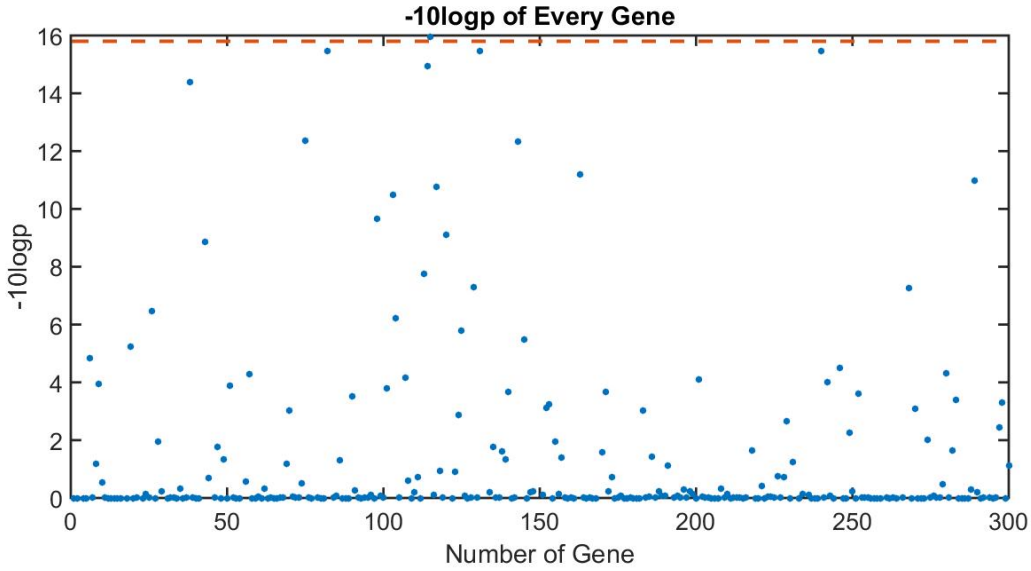

图 9　以布朗近似方法得到的 P-value 计算的 300 个基因的 $-10\log p$

5.4　本章小结

本文采用费舍尔方法对致病基因进行估计，分别对由费舍尔精确检验得到的 P-value 与 logistic 回归方法得到的 P-value 进行重要性分析，二者均得到四个致病基因，且其中有三个是重合的。由于费舍尔方法没有考虑位点之间的依赖关系，我们又采用了布朗近似对数据进行了分析，得到的致病基因与费舍尔方法得到的大体一致，但可减小 false positive rate。最终我们得到 102 号基因致病几率最大，55、217、265、293 号基因也有致病嫌疑。

6 问题四模型建立与求解

在实际的研究中，科研人员往往把相关的性状或疾病看成一个整体，然后来探寻与它们相关的位点或基因。现已知 10 种相关联的疾病的 1000 个样本的位点信息，根据这些信息确定可能的致病位点。为此本文采用改进的典型关联分析（canonical correlation analysis）方法对 10 种疾病与 9445 个位点这两组数据进行关联分析。

6.1 典型关联分析

在线性回归中，我们使用直线来拟合样本点，寻找 n 维特征向量 X 和输出结果（或者叫做 label）Y 之间的线性关系。其中 $X \in R^n$，$Y \in R$。然而当 Y 也是多维时，或者说 Y 也有多个特征时，我们希望分析出 X 和 Y 的关系。当然我们仍然可以使用回归的方法来分析，做法如下：假设 $X \in R^n$，$Y \in R^m$，那么可以建立等式（15）。

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{15}$$

其中 $y_i = w_i^T x$，形式和线性回归一样，需要训练 m 次得到 m 个 w_i。这样做的一个缺点是，Y 中的每个特征都与 X 的所有特征关联，Y 中的特征之间没有什么联系。我们想换一种思路来看这个问题，如果将 X 和 Y 都看成整体，考察这两个整体之间的关系。我们将整体表示成 X 和 Y 各自特征间的线性组合，也就是考察 $a^T x$ 和 $b^T y$ 之间的关系。这样的应用其实很多，举个简单的例子。我们想考察一个人解题能力 X（解题速度 x_1，解题正确率 x_2）与他/她的阅读能力 Y（阅读速度 y_1，理解程度 y_2）之间的关系，那么形式化为 $u = a_1 x_1 + a_2 x_2$ 和 $v = b_1 y_1 + b_2 y_2$，然后利用 Pearson 相关系数（16）来度量 u 和 v 的关系。

$$\rho_{u,v} = corr(u, v) = \frac{cov(u, v)}{\sigma_u \sigma_v} = \frac{E[(u - \mu_u)(v - \mu_v)]}{\sigma_u \sigma_v} \tag{16}$$

我们期望寻求一组最优的解 a 和 b，使得 $Corr(u, v)$ 最大，这样得到的 a 和 b 就是使得 u 和 v 就有最大关联的权重。求解 a、b 为一个优化问题，构造如公式（6.1）的优化问题

$$\begin{gathered} \max a^T \Sigma_{12} b \\ Subject \quad to \quad a^T \Sigma_{11} a = 1 \\ b^T \Sigma_{22} b = 1 \end{gathered} \tag{17}$$

其中，Σ_{11} 与 Σ_{22} 分别为 X 与 Y 的方差，Σ_{12} 是它们之间的协方差。求解方法是构造 Lagrangian 等式（18）。

$$\mathcal{L} = a^T \Sigma_{12} b - \frac{\lambda}{2}(a^T \Sigma_{11} a - 1) - \frac{\theta}{2}(b^T \Sigma_{22} b - 1) \tag{18}$$

由于篇幅限制，我们在这里不做求解的推导。

6.2　基于改进关联分析的致病位点研究

我们的目标是根据给出的 1000 个样本的 10 种疾病性状从 9445 个位点中找出致病位点，位点之间是有相互关联的，而且 10 种疾病之间也是有相互关联的，这种关联不一定是线性关联，而且我们的位点编码方式为 0、1、2 三个整数值所承载的信息有限。基于 [7]，我们在这里利用带惩罚的 CCA 并用使用最优缩放来解决这一问题。惩罚 CCA[8] 的算法6.1：

算法 6.1 惩罚关联分析

1. 标准化 Y 和 X.

2. 设置 k← 0.

3. 任意设置初始值 $\hat{u}^1$ 与 $\hat{v}^1$。例如设置 $\hat{u}^1 \leftarrow x_r$，$\hat{v}^1 \leftarrow y_s$，使得 $|cor(x_r, y_s)| = \max(|cor(x_j, y_i)|)$，其中 $r \in (1, \ldots, n)$，$s \in (1, \ldots, m)$，$j = 1, \ldots, n$ 与 $i = 1, \ldots, m$。

4. 如下迭代计算 u，v，a，b:

Repeat

(a) $k \leftarrow k + 1$

(b) $\hat{u}^k \leftarrow Y\hat{a}^{k-1}$，$\hat{v}^k \leftarrow X\hat{b}^{k-1}$（其中 $\hat{v}^1$ 与 $\hat{u}^1$ 由第三步给出）。

(c) 最小化 $\hat{u}^k$ 与 X 之间的距离得到变换矩阵 X^*

$$\tilde{c}_j = (G_j^T G_j)^{-1} G_j^T (\hat{u}^k)$$

其中 G_j 是以变量 j 和 g_j 为自变量的 $n \times g_j$ 的指示矩阵，g_j 为变量 j 的种类。限定 $\tilde{c}_j$ 得到 c_j^*。因此 $x_j^* = G_j c_j^*$，标准化 X^*。

(d) 利用 UST 计算 $\hat{a}^k$ 和 $\hat{b}^k$。

$$\hat{a}_i^k = (|\hat{v}^{kT} y_i| - \frac{\lambda_Y}{2})_+ sign(\hat{v}^{kT} y_i)$$

$$\hat{b}_j^k = (|\hat{u}^{kT} x_j^*| - \frac{\lambda_X}{2})_+ sign(\hat{u}^k x_j^*)$$

(e) 标准化 $\hat{a}_k$ 和 $\hat{b}^k$ 直到 $\hat{a}_k$ 和 $\hat{b}^k$ 收敛。

算法6.1中的 x_j^* 是通过最优缩放得到的。通过最优缩放可将离散型的位点数值编码转换成连续型的变量。换言之，数值编码 x_j 通过一种满足位点性质的变换（公式19）转换成连续型变量 x_j^*

$$x_j^* = \mathfrak{J}_j[x_j] \tag{19}$$

其中 $\mathfrak{J}_j$ 是 j 的测量转换，并且受到测量尺度的约束。

每个位点有三种编码方式：（1）两个主脱氧核苷酸 2，（2）一个主脱氧核苷酸一个次脱氧核苷酸，（3）两个次脱氧核苷酸 0。这些编码方式会有附加的、决定性的、隐性的或者无效的。这个信息决定了相应的转换变量的顺序：

$$\mathfrak{J}_j : (x_{aj} < x_{bj} < x_{cj}) \leftarrow \begin{cases} (x_{aj}^* \le x_{bj}^* \le x_{cj}^*), & or \\ x_{aj}^* \ge x_{bj}^* \ge x_{cj}^*, & . \end{cases} \tag{20}$$

其中 $a:2$，$b:1$，$c:0$。公式（20）表明杂合形式的位点效果介于其他二者之间。可能的变换如下：

- 附加效果 $\begin{cases}(x^*_{aj} < x^*_{bj} < x^*_{cj}), & or\\ (x^*_{aj} > x^*_{bj} > x^*_{cj}), & .\end{cases}$
- 隐性效果 $\begin{cases}(x^*_{aj} > x^*_{bj} = x^*_{cj}), & or\\ (x^*_{aj} = x^*_{bj} < x^*_{cj}), & .\end{cases}$
- 决定效果 $\begin{cases}(x^*_{aj} < x^*_{bj} = x^*_{cj}), & or\\ (x^*_{aj} = x^*_{bj} > x^*_{cj}), & .\end{cases}$
- 恒定效果 $(x^*_{aj} = x^*_{bj} = x^*_{cj})$

有了这些约束，CCA 中的最优缩放计算方法如下。一个 $n \times q$ 矩阵 X，包含 q 个位点信息，这些样本信息是从 n 个个体采集而来的并且其响应变量为 $\hat{u}$。令 G_j 为变量 j 的 $n \times g_j$ 的指示矩阵。其中 g_j 为变量 j 的种类总数。c_j 是变量 j 的分类量化结果。进而最优缩放如算法（6.2）所示。

算法 6.2 最优缩放

1. 标准化数据，使得平均值 $(G_j c_j) = 0$，$c_j^T G_j^T G_j c_j = n$。并且平均值 $(\hat{u}) = 0$，$\hat{u}^T \hat{u} = n$。

2. 设置初始值 $v_j = cor(\hat{u}, x_j)$，$z = \sum_{j=1}^{q} v_j G_j c_j$。

3. 将 X 转换为 X^*:

   For $j = 1, \ldots, q$

   (a) 从 z 中移除变量 j：$z_j = z - v_j G_j c_j$

   (b) 计算 c_j 的非约束变换：$\tilde{c}_j = (G_j^T G_j)^{-1} G_j^T (\hat{u} - z_j)$

   (c) 约束并标准化 $\tilde{c}_j$ 得到 c_j^*

   (d) 更新 v_j：$v_j^* = n^{-1} \tilde{c}_j^T (G_j^T G_j) c_j^*$

   (e) 更新 z：$z = z_j + v_j^* G_j c_j^*$

   直到 v 收敛。

基于惩罚 CCA，我们得到了权重向量，并把权重向量进行归一化，以 10 为权重的最大值，得到的权重图如图 10 所示。由图可知，“rs351617”位点的致病可能性最大，“rs7538876”位点和“rs4949516”位点致病可能次之，“rs780983”位点、“rs12746773”位点、“rs406985”位点和“rs716325”位点也有致病的可能。

6.3 本章小结

针对多性状多位点的相互关联问题，我们采用改进关联分析方法对两个集合的关联性进行了分析。传统的关联分析只能解决线性关联。为此我们使用了惩罚典型相关分析，并且通过最优缩放将位点的离散数值编码转换为连续变量，便于进行关联计算。结果显示有 7 个位点治病可能性较大，另外有 23 个位点也有致病的可能性。

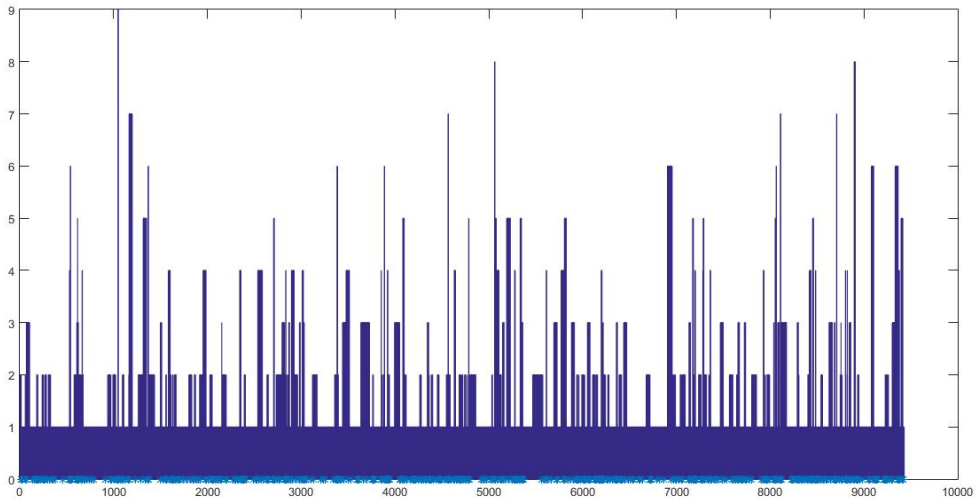
9
8
7
6
5
4
3
2
1
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

图 10　各位点的权重图

7 总结

本文主要研究了三个问题。在对位点类型进行 0、1、2 数值编码的基础上，我们首先对致病位点进行了研究。采用的方法为假设检验的方法，并用三种不同的检验方法分别对数据进行了处理，即费舍尔精确检验、logistic 回归检验和 Cochran-Armitage trend 检验。三种方法得到的结果基本一致，即“rs2273298”位点的致病可能性最大。在对三百个含不同位点的基因进行致病性研究的时候，我们首先采用了费舍尔方法对每个基因的致病概率进行分析。由于费舍尔方法假定位点间是相互独立的，这就使得位点间可能存在相互依赖关系被忽略了，可能会造成结果的不准确。为此，我们又采用了布朗方法定位致病基因，结果显示布朗方法与费舍尔方法得到的结果非常相近。而且我们也分别用上一问题的三种检验方法得到的 P-value 来进行布朗方法与费舍尔方法的计算，结果也趋于一致。结果表明 102 号基因的致病概率最大。

在对 10 种相关联的疾病的致病位点定位问题中，我们使用典型关联分析（CCA）方法。由于传统 CCA 方法只能处理线性关联，而且由于位点的数值编码仅为 0、1、2 三个数值，信息量较少，直接采用基本的 CCA 方法会带来较大误差甚至找到不正确的位点（即 false positive)。因此本文中使用了带最优缩放的惩罚 CCA 方法，最优缩放可将位点的数值编码转换为连续变量，惩罚 CCA 考虑了每个变量集合内部不同元素之间的非线性相关关系。结果表明“rs351617”位点的致病可能性最大。由于本文的模型都是基于假设检验以及关联分析这两种统计学方法的，因此结果给出的都是位点或基因的致病概率，使用不同的接受值得到的致病位点或基因的数量可能有所不同。而且由于目前还没有一个统一的、公认的最优接受值，因此模型所给出的结果具有一定的不确定性。在将来的工作及研究中，我们考虑首先根据已知信息进行数据分析，这可加强问题的约束。增加更多的紧约束，可降低结果的不确定性。

参考文献

[1] High-Seng Chai et al. “GLOSSI: a method to assess the association of genetic loci-sets with complex diseases”. In: *BMC bioinformatics* 10.1 (2009), p. 1.

[2] Graham JG Upton. “Fisher’s exact test”. In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* (1992), pp. 395–402.

[3] Michael C Whitlock. “Combining probability from independent tests: the weighted Z-method is superior to Fisher’s approach”. In: *Journal of evolutionary biology* 18.5 (2005), pp. 1368–1373.

[4] Ramon C Littell and J Leroy Folks. “Asymptotic optimality of Fisher’s method of combining independent tests”. In: *Journal of the American Statistical Association* 66.336 (1971), pp. 802–806.

[5] Ramon C Littell and J Leroy Folks. “Asymptotic optimality of Fisher’s method of combining independent tests II”. In: *Journal of the American Statistical Association* 68.341 (1973), pp. 193–194.

[6] Morton B Brown. “400: A method for combining non-independent, one-sided tests of significance”. In: *Biometrics* (1975), pp. 987–992.

[7] Sandra Waaijenborg and Aeilko H Zwinderman. “Correlating multiple SNPs and multiple disease phenotypes: Penalized nonlinear canonical correlation analysis”. In: *Bioinformatics* (2009), btp491.

[8] Sandra Waaijenborg, Philip C Verselewel de Witt Hamer, and Aeilko H Zwinderman. “Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis”. In: *Statistical Applications in Genetics and Molecular Biology* 7.1 (2008).

[9] Chen Suo et al. “Analysis of multiple phenotypes in genome-wide genetic mapping studies”. In: *BMC bioinformatics* 14.1 (2013), p. 1.